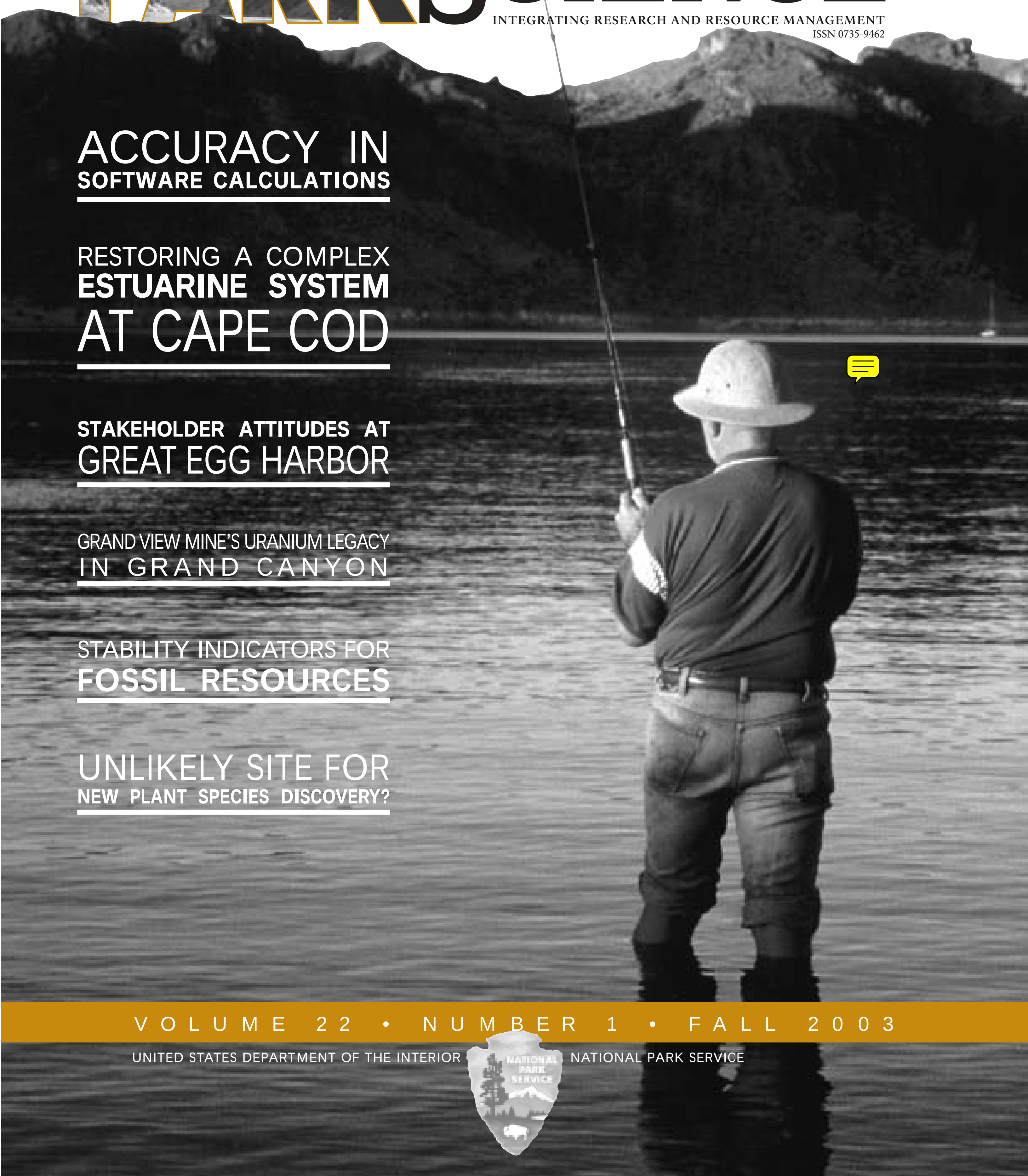
PARK SCIENCE

INTEGRATING RESEARCH AND RESOURCE MANAGEMENT
ISSN 0735-9462



ACCURACY IN SOFTWARE CALCULATIONS

RESTORING A COMPLEX ESTUARINE SYSTEM AT CAPE COD

STAKEHOLDER ATTITUDES AT GREAT EGG HARBOR

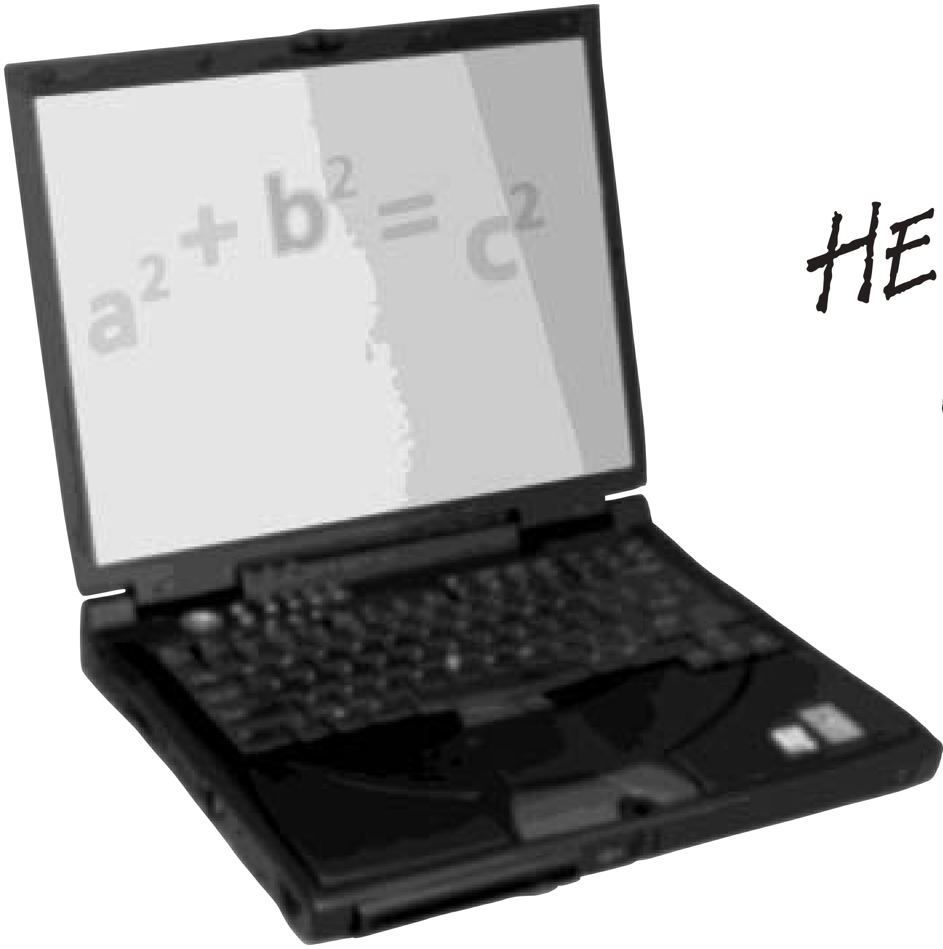
GRAND VIEW MINE'S URANIUM LEGACY IN GRAND CANYON

STABILITY INDICATORS FOR FOSSIL RESOURCES

UNLIKELY SITE FOR NEW PLANT SPECIES DISCOVERY?

VOLUME 22 • NUMBER 1 • FALL 2003

UNITED STATES DEPARTMENT OF THE INTERIOR NATIONAL PARK SERVICE



HEADS UP!

Uncertainty in software accuracy

By Roy Irwin

Disclaimer: No government endorsement is implied for software examples mentioned. No representation is made that all worthy software alternatives are mentioned or that all potential kinds of problems are described.

Several recent publications have discussed potential problems with accuracy of statistical calculations in MS Excel, including Cox (2000), Cryer (2001), Knusel (1998), McCullough (1998), McCullough (1999), and McCullough and Wilson (1999). Considering the prevalent use of MS Excel by employees of the National Park Service (NPS), what should NPS scientists know about the findings of these publications and the advice of statistical experts? Are calculations in MS Excel—the "standard spreadsheet" used by NPS employees—accurate enough to be credible? Should the National Park Service heed the strong condemnations of Cryer (2001)—"Friends Don't Let Friends Use Excel for Statistics!"—or the more subdued advice of McBride (see details below) that "MS Excel is fine for basic calculations, except percentiles and odd data sets"?

After reviewing the findings in these papers and discussing the issues with independent experts in statistics, I have decided that the answer depends on the situation; therefore, it is difficult to provide blanket "one size fits all" guidance. Instead, I will attempt to alert the National Park Service to some of the issues and summarize some of the basic information that NPS scientists should know.

What's "accurate enough"?

First, I need to consider briefly how to define "accurate enough" in the context of software calculations. Uncertainty in accuracy of software is just one of many known sources of uncertainty. Another source is measurement uncertainty. For environmental studies, measurement uncertainty (factoring in both precision and systematic error, i.e., bias) is seldom lower than plus or minus 3% and is often much higher for parameters like pesticides or for observations such as "percent embeddedness in cobbles" in stream bottom sediments.

"Uncertainty in accuracy of software is just one of many known sources of uncertainty."

The National Institute of Standards and Technology (NIST) publishes International Organization for Standardization or ISO-compatible sum-of-squares equations for combining uncertainty from many sources (Taylor and Kuyatt 1994). In these equations, if the (standard deviation) contributor to uncertainty from one source is five times lower than another contributor, it is considered trivial and, therefore, not considered in the overall uncertainty equation.

In some cases where the MS Excel answer is not exactly right, that contribution to uncertainty may be considered trivial compared to others. For example, in one of the publications outlining deficiencies in MS Excel (McCullough and Wilson 1999), a standard deviation calculated by MS Excel was 0.0790105482336451. This was compared to a "correct" standard deviation of 0.0790105478190518 calculated by a benchmark standard.

Now What Was That Rounding Rule?

Is 0.0790105482336451

really different from **0.0790105478190518**

or are they both **0.079**?

Although environmental specialists have to be careful not to round numbers too aggressively before using them in subsequent calculations, typically neither rounding rules nor uncertainty estimates justify using more than two or three significant figures in final environmental measurement results. Rounding rules for final results often amount to a crude (better than nothing) way to account for uncertainty. More modern ways to account for uncertainty include confidence intervals for summary statistics and estimates of measurement uncertainty for individual observations. Such uncertainty estimates for environmental data sets often make it clear that we are seldom certain of more than three significant figures, and sometimes we are not certain of more than one or two. The difference between the two examples of standard deviation values shown in the illustration on this page would typically not approach one-fifth of the total estimated uncertainty from combined sources, and thus would usually be considered trivial and not added to overall uncertainty equations.

Furthermore, neither measurement uncertainty nor uncertainty in summary statistics usually account for the largest contributions to overall uncertainty. Other common contributors to uncertainty include model uncertainty, uncertainty in how representative the sample is of the larger population, and errors arising from not using software correctly. In fact, using software incorrectly may be more common when using dedicated statistical software than when using the more familiar, "user-friendly," and ubiquitous MS Excel. Other common sources of errors and uncertainty include choosing the wrong analysis, not meeting critical assumptions, and cumulative rounding errors. Furthermore, there are inherent uncertainties related to our imperfect knowledge of biology and physical science, wrong or crude theories, and various sampling errors. Such errors typically vary in magnitude in both time and space. Collectively, these "additional" sources of uncertainty are probably of greater magnitude than software calculation errors, particularly for simple summary statistics.

I have as yet to see any examples where MS Excel gave an answer for simple statistics—such as a mean or a sample standard deviation—"different enough" not to be considered trivial in comparison with other sources of uncertainty in environmental variables. Many users conclude that for very simple calculations—like a population standard deviation, a mean, or even a 95% "t distribution" confidence interval—answers in MS Excel may be "accurate enough" for analyzing environmental data sets and for commonly used statistics. Hence, given that NPS offices typically already have MS Excel, investment in dedicated statistical software may be difficult to justify.

Graham McBride (National Institute of Water and Atmospheric Research in New Zealand) uses and deems MS Excel useful for some routine analyses, for less-routine multiple one-sided (TOST) tests for inequivalence, and for some Bayesian statistics that use routine functions. McBride points out that MS Excel is fine for basic calculations such as means, standard deviations, and t values, and that most of the criticisms of MS Excel are valid only for very odd data. Although MS Excel may have problems with percentiles, so does the software S-Plus. Except for SAS, statistics packages typically do not explain that there is no one "right" way to calculate percentiles, let alone tell you which one they use. Users should not assume that a particular computer package is automatically acceptable for their purposes; they should seek expert statistical advice when needed. The level of explanation in manuals and help files is often poor (Graham McBride, personal communication, 2002).

"Users should not assume that a particular computer package is automatically acceptable for their purposes; they should seek expert statistical advice when needed."

"Calculating complicated statistics ... may be risky when using MS Excel."

Calculating complicated statistics (e.g., regression statistics) may be risky when using MS Excel. Such calculations performed in MS Excel may not yield acceptable levels of accuracy. Hence, for complicated statistics, the investment in dedicated statistical software would be more easily defended, especially in legally or professionally contentious settings. Users should keep in mind, however, that even dedicated statistics software packages can have difficulty with complicated data sets: those that involve large numbers (usually greater than six digits), have a very large sample size (high n), or have constant leading digits (e.g., 90000001, 90000002, and 90000003).

Odd or difficult data sets are typically rare in environmental work, but this may change as more and more continuous data readout probes are used and the National Park Service accumulates long-term data sets from Vital Signs and other long-term monitoring programs.

What sample sizes are too large?

If the individual numbers are high six digits, and the sample size is anything but small, the data set may be starting to "become difficult" even for the relatively simple standard deviation determinations in MS Excel. However, determining the case-by-case limitations may be challenging for ordinary users. It would be nice if the software makers explained particular conditions for each software-hardware combination: if the sample size is less than "n" and no numbers have more than "z" digits, then large sample or large number-related problems will not arise, but they do not (Bruce McCullough, Federal Communications Commission, personal communication, 2000).

In cases where percentiles or complicated statistics are calculated in MS Excel, I recommend that for the purpose of quality assurance, data analyses should be replicated on at least one "dedicated" statistical software program to help ensure accuracy. This software should be dedicated to statistical tasks, such as the following (not necessarily complete) list of typical or widely used examples: SAS, SPSS, SYSTAT, MATHEMATICA, EquivTest, WQStat, MINITAB, STATGRAPHICS, STATA, MAPLE, or S-PLUS.

Words of caution

When complicated statistics are to be calculated, when very large or otherwise difficult data sets are to be analyzed, or when legal or rigorous professional challenges are expected, a logical first step for "additional" quality assurance in software accuracy would be to compare results of analyses of published standard data sets of certified NIST correct answers for calculations of the same type as will be performed for the work at hand (see NIST 2000).

"The more complicated the statistic, the more can go wrong."

One thing users should keep in mind: the more complicated the statistic, the more can go wrong. Therefore, we should not be surprised to see more software errors of multivariate or other complicated procedures than in calculations of the sample mean. McCullough's 1999 summary confirms more errors on the relatively complicated statistical procedures (such as multi-factor ANOVAs and nonlinear regressions) than on the relatively simple univariate procedures (McCullough 1999).

Cox (2000) likewise suggested that areas in which MS Excel may be unreliable include some relatively complex tasks and unusual data sets:

- Standard deviations and statistics (e.g., t-tests) relying on standard deviation calculations that have large numbers with low variation
- Multiple regressions with very high collinearity
- Nonlinear regression problems
- Distribution tail areas beyond about 10-6
- Procedures (e.g., bootstrap) that rely on a good random number generator

Cryer (2001) offers one of the more strongly worded cautions against using MS Excel for statistics, citing not only problems with regression analyses but also with graphing functions and even with the algorithms used to calculate simple summary statistics such as a standard deviation. Cryer points out that MS Excel has problems with how it treats missing data and same values; it also, among other things, may display many more digits than are warranted.

The problem of displaying too many digits is not limited to MS Excel, however. Some programs allow users to specify the number of digits reported, and in any case, it is typically up to users to apply logical rounding rules, such as those summarized by Irwin (2003). In trials for calculating a sample standard deviation with MS Excel and other dedicated statistical software (i.e., SYSTAT), both programs usually tended to return the same standard deviation when rounded to a reasonable number of digits, so I remain unconvinced that there is a major problem with the way MS Excel calculates sample standard deviations for typical environmental data sets.

Reliability of other software programs

What about reliability of other software programs? Although potential problems with MS Excel have been well publicized, it is less well known that other software and software-hardware combinations can have problems and limitations too. Even dedicated statistical software programs can have difficulty with very difficult or contrived data sets. Using relatively difficult tests in 1999, flaws were discovered in the then-current versions of widely used statistical packages such as SAS, SPSS, and S-Plus (McCullough 1999). In a separate comparison involving relatively simple tests, problems were noted with some of the dedicated statistical packages. In general, however, more serious problems were reported with MS Excel (Landwehr and Tasker 1999).

The picture gets further muddied when considering various multivariate, ordination, and phylogenetic classification programs. For example, taxonomists use such programs to help classify the phylogenetic relationships between species. Many schemes are used, and some programs are developed by individuals and have very little documentation. Users should be aware that the answers may be suspect and are not always consistent between different programs. For example, in one trial, taxonomists tried several popular multivariate software programs on identical data sets and came up with very different answers. When they entered data in a different order, they got yet other answers (Terry Frest, consultant and malacologist, personal communication, 2000).

Controlling a measurement process is difficult without controlling for systematic error (bias). Bias is difficult to estimate if one cannot identify what the right, or at least the "expected," answer is. This makes quantifying uncertainty in the answers from many multivariate, ordinations, and classification programs very difficult. Typically only expert opinion, nonstatistical, or qualitative estimates of uncertainty are possible.

For additional details on estimating measurement and simple types of model uncertainty, see Irwin (2003).

References

Cox, N. 2000. Use of Excel for statistical analysis. Available at <http://www.agresearch.cri.nz/Science/Statistics/exceluse.htm> (accessed 6 July2003).

Cryer, J. 2001. Problems with using Microsoft Excel for statistics. Joint Statistical Meetings, Atlanta, Georgia. Available at <http://www.stat.uiowa.edu/~jcryer/JSMTalk2001.pdf> (accessed 1 July 2003).

Irwin, R. J. 2003. Vital Signs long-term aquatic monitoring projects. Part B. Planning process steps: issues to consider and then to document in a detailed study plan that includes a quality assurance project plan (QAPP) and monitoring "protocols" (standard operating procedures). Available at <http://science.nature.nps.gov/im/monitor/protocols/wqPartB.doc> (accessed 21 October 2003).

Knusel, L. 1998. On the accuracy of statistical distributions in Microsoft Excel 97. Computational Statistics and Data Analysis 26:375–377. Available at <http://www.stat.uni-muenchen.de/~knuesel/elv/excelacc.pdf> (accessed 1 July 2003).

Landwehr, J. M., and G. D. Tasker. 1999. Notes on numerical reliability of several statistical analysis programs. Open-file report 99-95. U.S. Geological Survey, Reston, Virginia. Available at <http://water.usgs.gov/pubs/of/ofr99-95/ofr.99-95.pdf> (accessed 1 July 2003).

McCullough, B. D. 1998. Assessing the reliability of statistical software. Part I. The American Statistician 52:358–366. Available at <http://www.amstat.org/publications/tas/mccull-1.pdf> (accessed 1 July 2003).

McCullough, B.D. 1999. Assessing the reliability of statistical software. Part II. The American Statistician 53:149–159. Available at <http://www.amstat.org/publications/tas/mccull.pdf>(accessed 1 July 2003).

McCullough, B. D., and B. Wilson. 1999. On the accuracy of statistical procedures in Microsoft Excel 97. Computational Statistics and Data Analysis 31:27–37. Available at <http://www.seismo.unr.edu/ftp/pub/updates/louie/mccullough.pdf> (accessed 1 July 2003).

National Institute of Standards and Technology (NIST). 2000. Statistical reference datasets. NIST, Gaithersburg, Maryland. Available at <http://www.itl.nist.gov/div898/strd/> (accessed 1 July 2003).

Taylor, N., and C. E. Kuyatt. 1994. Guidelines for evaluating and expressing the uncertainty of NIST measurement results. Technical note 1297. National Institute of Standards and Technology, Gaithersburg, Maryland. Available at <http://physics.nist.gov/Document/tn1297.pdf> (accessed 1 July 2003).

About the Author

Roy Irwin is Senior Contaminants Specialist with the Water Resources Division in Fort Collins, Colorado. He can be reached at 970-225-3520 and roy_irwin@nps.gov. He wrote for *Park Science* in volume 16, number 2, on pseudoreplication, a complex issue in experimentation that can lead to invalid conclusions.